BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37184
doi: 10.3897/biss.3.37184

OPEN ACCESS

Conference Abstract

What's in this Collection Dataset? Semantic Annotation with GATE

Felicitas Löffler‡, Birgitta König-Ries‡

‡ Friedrich Schiller University Jena, Jena, Germany

Corresponding author: Felicitas Löffler (felicitas.loeffler@uni-jena.de), Birgitta König-Ries (birgitta.koenig-ries@uni-jena.de)

Received: 12 Jun 2019 | Published: 18 Jun 2019

Citation: Löffler F, König-Ries B (2019) What's in this Collection Dataset? Semantic Annotation with GATE. Biodiversity Information Science and Standards 3: e37184. https://doi.org/10.3897/biss.3.37184

Abstract

Semantic annotations of datasets are very useful to support quality assurance, discovery, interpretability, linking and integration of datasets. However, providing such annotations manually is often a time-consuming task . If the process is to be at least partially automated and still provide good semantic annotations, precise information extraction is needed. The recognition of entity names (e.g., person, organization, location) from textual resources is the first step before linking the identified term or phrase to other semantic resources such as concepts in ontologies. A multitude of tools and techniques have been developed for information extraction. One of the big players is the text mining framework GATE (Cunningham et al. 2013) that supports annotation rules, semantic techniques and machine learning approaches. We will run GATE's default ANNIE pipeline on collection datasets to automatically detect persons, locations and time. We will also present extensions to extract organisms (Naderi et al. 2011), environmental terms, data parameters and biological processes and how to link them to ontologies and LOD resources, e.g., DBPedia (Sateli and Witte 2015). We would like to discuss the results with the conference participants and welcome comments and feedbacks on the current solution. The audience is also welcome to provide their own datasets in preparation for this session.

© Löffler F, König-Ries B. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords

semantic annotation, collection data, GATE

Presenting author

Felicitas Löffler

Presented at

Biodiversity_Next 2019

References

- Cunningham H, Tablan V, Roberts A, Bontcheva K (2013) Getting more out of biomedical documents with GATE's Full Lifecycle Open Source Text Analytics. PloS Computational Biology 9 (2): e1002854. https://doi.org/10.1371/journal.pcbi.1002854
- Naderi N, Kappler T, Baker CJO, Witte R (2011) OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. Bioinformatics 27 (9): 2721-2729. https://doi.org/10.1093/bioinformatics/btr452
- Sateli B, Witte R (2015) Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud. PeerJ Computer Science 1: e371. https://doi.org/10.7717/peerj-cs.37